Comparison of different test metrics for detecting single-cell RNA-seq batch effects

Amirhossein Alvandi, Flore Uzan and Yuting Xu

December 4th, 2020

Abstract

In molecular biology, we use the term batch effect to refer to any heterogeneity due to experimental factors or non-biological factors in an experiment that causes changes in the data. Several strategies has been introduced in order to remove or reduce the batch effects. Batch correction methods are often evaluated by visual inspection but remains imprecise. One of the common methods to quantify the extent to which they remove batch effects while preserving biological variability is called the kBET method. We propose to revisit the kBET method by comparing it with different test metrics on different datasets from mild to strong batch effects.

1 Introduction

Technological advances in the recent years have increased our ability to generate high-throughput single-cell gene expression data. Single-cell data is often compiled from multiple experiments with differences in capturing times, handling personnel, reagent lots, equipments, and even technology platforms. These differences lead to large variations or batch effects in the data, and can confound biological variations of interest during data integration. Batch effects can be highly nonlinear, making it difficult to correctly align different datasets while preserving key biological variations.

As such, effective batch-effect removal is essential. Various methods have been proposed in literature to remove or reduce the cell- bias and batch-effects in single-cell data while preserving biological variability: linear regression such as Combat, Seurat's canonical correlation analysis and projection of mutual nearest neighbors.

In order to analyze the batch effect correction, visualization methods including clustering or principal component analysis have been proposed. However, they are not suited for very large number of samples and they lack precision.

One of the most renowned methods that quantifies the batch effect in scRNA-seq data is kBET method (Section 2). It measures if a replicated experiment is well-mixed, i.e., it ensure that any subset of neighboring samples has the same distribution of batch labels as the full dataset. Figure 1, illustrates a dataset with both sources of variation including biological and technical variations. In the left panel, we can see that the data is well-mixed since the proportions of different batch labels in any neighborhood is similar to the global distribution. However, in the right panel, we can observe that the data points from respective batches tend to cluster with their 'peers', and batch label proportions differ considerably between arbitrarily chosen neighborhood.

Conditions of validity of parametric tests are different than for non-parametric tests and the decision of which statistical test to use depends on the research design, the distribution of the data, and the type of variable. Parametric tests make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the assumption that the population data are normally distributed. Non-parametric tests are "distribution-free" and, as such, can be

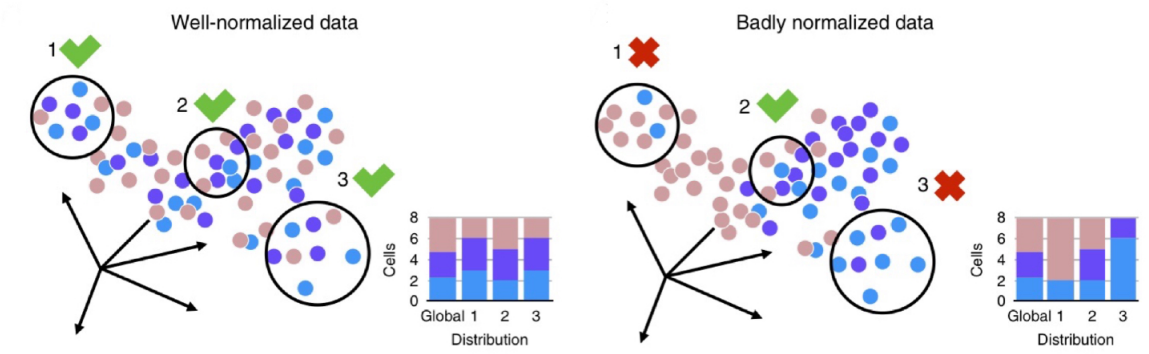

Figure 1: On the left, the data is well-mixed since the proportions of different batch labels in any neighborhood is similar to the global distribution. On the right, the data points from respective batches tend to cluster with their 'peers', and batch label proportions differ considerably between arbitrarily chosen neighborhood.

used for non-Normal variables. Here, we introduce some possible alternative test metrics in the kBET method. We keep the same structure of the test, explained in Section 2, however, we convert the χ^2-based test with three alternative nonparametric tests introduced in Section 3. We compare these test metrics that allow us to diagnose and quantify any potential batch effects that are present in scRNA-seq datasets in order to define the most appropriated test.

2 kBET method

Single-cell transcriptomics is a versatile tool for exploring heterogeneous cell populations, but as with all genomics experiments, batch effects can hamper data integration and interpretation. The success of batch-effect correction is often evaluated by visual inspection of low-dimensional embeddings, which are inherently imprecise. This method is a user-friendly, robust, and sensitive k-nearest-neighbor batch-effect test. In this section, first, we introduce the methodology. We then elaborate on the limitations of kBET, which motivated us to alter the algorithm by using three different nonparametric tests.

2.1 Algorithm

kBET uses a χ^2-based test for random neighborhoods of fixed size to determine whether they are well mixed, followed by the averaging of the binary test results to return an overall rejection rate.

We consider the full gene expression dataset D $= \{x_1, ..., x_n\}$ with $x_j \in R^g$ and X $\in R^{n \times g}$ which is the corresponding data matrix with n samples and g genes.

For each batch variable, we consider l categories such that n_i denotes the number of samples in batch i. We define $\mathrm{f}_i = \frac{n_i}{n}$ is the *global* fraction of sample in batch i and $\tilde{f}_i$ is the *local* fraction of sample in batch i in some subset N $\subset$ D. In particular, we consider k nearest neighbors.

Let n_{ji}^k denote the number of cells in batch i that are in subset j of size k.

The null hypothesis can be formulated such that:

H_0: $\tilde{f}_i = \mathrm{f}_i \; \forall$ i $\in \{1, ..., l\} \; \forall$ subsets N $\subset$ D.

Intuitively, a replicated experiment is well mixed if a subset of neighboring samples has the same distribution of batch labels as the full dataset, i.e., the local distribution is equal to the global distribution.

To test the null hypothesis, we define a neighborhood subset $\mathrm{N}_j = \mathrm{x}_j \cup \{ x_s \mid \mathrm{s}$ among k-1 nearest neighbors of $j\}$. We compute the k-nearest Neighbors with the cover-tree algorithm (FNN R package). Finally, we compute the first 50 largest eigen values with the svd function and we use the reduce dataset to find nearest neighbors.

We first test the null hypothesis in each subset N_j of a given sequence of subsets. In other words, we test whether the distribution of n_{ji}^k with respect to i matches the distribution under the null

hypothesis. Then, we compute the average rejection of rate S over all tests - a test statistic for the whole dataset.

The k-BET method uses the Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution. It evaluates whether an observed frequency distribution differs from a theoretical distribution. The formula of the Pearson's cumulative test is given below:

$$\tilde{\chi}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where O_i is the number of observations of class i, N is total number of observations, $E_i = Np_i$ is the expected (theoretical) count of batch i, asserted by the null hypothesis that the fraction of class i in the population is p_i and $n =$ the number of classes.

In our context, we consider that n_{ji}^k is Gaussian- distributed with respect to i and we define the Pearson χ^2 test:

$$k_j^k = \sum_{i=1}^{l} \frac{(n_{ji}^k - f_i x k)^2}{f_i x k} \sim \chi_{l-1}^2$$

where n_{ji}^k is the number of cells in batch i that are in subset j of size k, $\mathrm{f}_i = \frac{n_i}{n}$ is the *global* fraction of sample in batch i, and χ_{l-1}^2 denotes χ^2 distribution with $l-1$ degrees of freedom.

The p-value for each k_j^k is $1 - F_{l-1}(k_j^k)$ where $F_{l-1}(x)$ denotes the cumulative distribution function of the χ^2 distribution of the $l-1$ degrees of freedom. A low rejection rate means that the data is well-mixed.

2.2 Limitations

Even though parametric tests usually have greater statistical power than non-parametric tests, they require strong assumptions. In particular, here, we consider that n_{ji}^k is Gaussian-distributed with respect to i in order to use the Pearson's χ^2 test. Like any parametric tests, we assume the underlying statistical distributions in the data.

We commonly decide to use a parametric test if the mean accurately represents the center of the distribution and the sample size is large enough. However, if the median better represents the center of the distribution, we consider the non-parametric test even with a large sample.

It might be better for our study area to assess the median since we our counting and comparing the number of each labels of the local distribution with the global distribution. The mean is not always the better measure of central tendency for a sample. Even though you can perform a valid parametric analysis on skewed data, that doesn't necessarily equate to being the better indicator.

Besides, non-parametric tests are valid when our sample size is small and your data are potentially nonnormal. With the χ^2 test, no more than 20% of the expected counts could be less than 5 and all individual expected counts are 1 or greater.

Finally, we observe that the Pearson's chi-square is far too conservative. The actual Type I error rate is well below the nominal alpha. We want to compare the three non-parametric tests and observe if the batch effect correction is better represented by the tests.

3 Alternative test metrics

In order to achieve a comprehensive understanding of the possible drawbacks of using chi-squared test, here we discuss some alternative nonparametric test metrics specifically to evaluate their performance in situations where we do not have access to a massive sized dataset.

3.1 Mann Whitney U-Test

The Mann-Whitney U test is a nonparametric statistical significance test for determining whether two independent samples were drawn from a population with the same distribution. The strategy is to compare two populations rank ordering instead of comparing two populations means. By using the critical values tables, it is possible to assess the degree to which any observed difference is a result of chance or fluke.

In order to do so, we combined the two samples together and rank all observations in increasing order of magnitude, ignoring which group they come from. If two observations have the same magnitude, regardless of the group, then they are given an average ranking. The strategy

is to determine if the values from the two samples are randomly mixed in the rank ordering or if they are clustered at opposite ends when combined. A random rank order would mean that the two samples are not different, while a cluster of one sample values would indicate a difference between them.

The default assumption or null hypothesis is that there is no difference between the distributions of the data samples (H_0: The two populations have the same distributions). Rejection of this hypothesis suggests that there is likely some difference between the samples.

The Mann Whitney U test statistic obtained such that U_{obt} is the smaller of

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_1 and R_2 are the sums of the ranks in groups 1 and 2, n_1 and n_2 are the sample size of the groups 1 and 2 and U_1 and U_2 are the statistics of the groups 1 and 2.

We reject H_0 if $U_{obt} \leq U_{crt}$, where U_{crt} is determined by U_{obt}, n_1, n_2 and the level of significance α:

$$z_u = \frac{|U_{obt} - (\frac{n_1 n_2}{2})|}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

3.2 Wilcoxon T-Test

The Wilcoxon signed ranks test is a nonparametric statistical procedure that is used to compare two samples that are paired, or related. The parametric equivalent to the Wilcoxon signed ranks is associated to the t-test for matched pairs. Again, we compare the two populations ranks.

We start by compute the difference between the two samples. However, unlike the Mann Whitney U test, we ignore the sign of the difference and order rank the absolute value of the difference. We add together the ranks belonging to score with positive sign and do the same for the negative sign and finally add them together.

$$W = |\sum_{i=1}^{N_r} [sgn(x_{2,i} - x_{1,i}).R_i]|$$

which is the absolute value of the sum of the signed ranks, where N_r is the sample size excluding pairs that are equal and R_i denote the rank.

For N_r larger than 10, the sampling distribution of W converges to a normal distribution. Thus, for $N_r \geq 10$, a z-score can be calculated:

$$z = \frac{W - 0.5}{\sigma_W}$$

$$\sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$

as if $z > z_{critical}$ then reject H_0.

For $N_r < 10$, W is compared to a critical value from a reference table, therefore, we reject that the two medians are the same.

If $W_{critical,N_r}$, then reject H_0.

3.3 Kolmogorov–Smirnov Test

Two-sample Kolmogorov–Smirnov test (K-S test) is used to test whether two independent samples come from the same distribution. It compares empirical distribution functions of two samples and computes the maximum difference between them. A large maximum difference indicates a difference between the two sample distributions.

To perform the two-sample K-S test, we first sort the combined sample from the smallest to the largest and compute the empirical cumulative distribution function (CDF) of the two samples respectively using

$$F_n(x) = P(X < x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

Next, we compute the test statistic $D_{n,m}$, which is the maximum difference between two empirical CDFs, using

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Where n and m are the sizes of first and second sample respectively.

The null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n*m}}$$

Where $c(\alpha)$=$\sqrt{-ln(\frac{\alpha}{2} * \frac{1}{2})}$ in general.

Note that the two-sample K-S test checks whether the two data samples come from the same distribution but it does not specify what

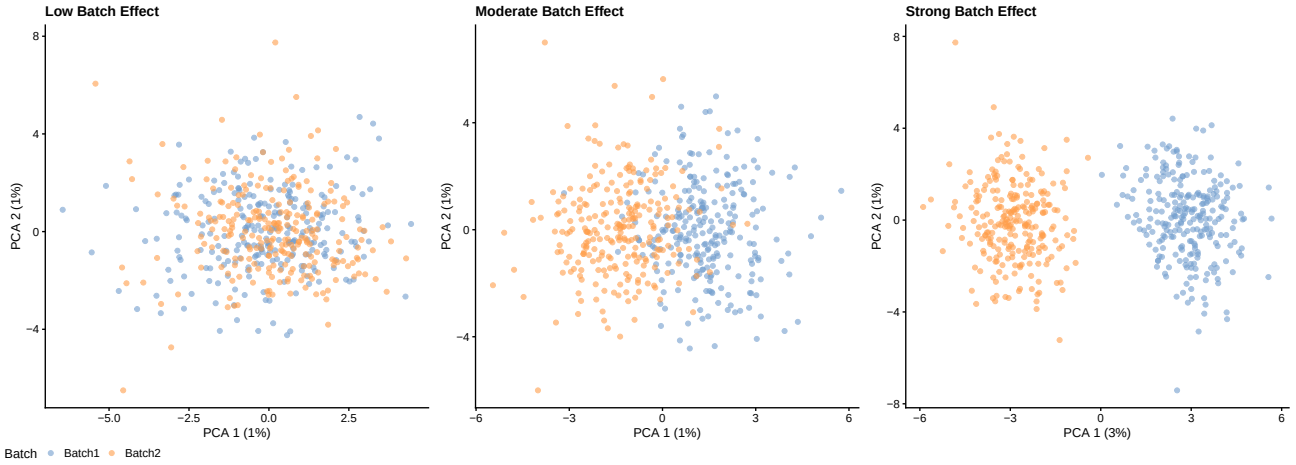

Figure 2: Simulated data with different strengths of batch effects. Low (location factor = 0.01, scale factor = 0.01); Moderate (location factor = 0.05, scale factor = 0.05); Strong (location factor = 0.1, scale factor = 0.1).

that common distribution is (e.g. whether it's normal or not normal). A shortcoming of the Kolmogorov–Smirnov test is that it is not very powerful because it is devised to be sensitive against all possible types of differences between two distribution functions.

4 Data Simulation

Here, we used two different approaches for simulating scRNA-seq data and batch effects. The first method samples the data points from a negative binomial model, where the second approach is based on the Splat package in R.

4.1 Zero-inflated negative binomial model

The first model that we used is based on a zero-inflated negative binomial distribution for count data where mean expression levels for each gene were sampled from a scaled beta-distribution

$$\mu \sim Beta(a, b).c$$

with parameters $a = 2$, $b = 5$ and $c = 100$. In addition, the dropout probability for each simulated gene $j \in \{1, \ldots, G\}$ in batch $i \in \{1, 2\}$ is modeled by the logistic (sigmoid) function

$$p_{ij} = sigm(-(\beta_0 + \beta_{1,i}\mu_{ij}))$$

where we chose $\beta_0 = -1.5$ and $\beta_{1.i} = \frac{1}{med(\mu_i)}$. We must note that here, every single sample was drawn from $s_{ij} \sim NB(\mu_{ij}, \theta|Ber(p_{ij}))$.

Regarding the batch effect strength with the parameters of the first batch set up, the mean expression level of the second batch μ_2 are subject to different degrees of variation. We multiply 1%, 10%, and 20% of the mean gene expression levels as follows

$$\mu_{2,j} = \begin{cases} \mu_{1,j}.\gamma & where \quad j \in \{1, \ldots, h.G\} \\ \mu_{1,j} & otherwise \end{cases}$$

where $\gamma \sim Gamma(1, 1)$, and $h \in \{0.01, 0.1, 0.2\}$. The gamma distribution parameters have been selected such that the expected value of the sampled mean expression levels remains the same.

4.2 Simulating single-cell RNAseq count data using Splat

The core of the Splat model is a gamma-Poisson distribution used to generate a gene by cell matrix of counts. Gene mean expression μ_i is modelled by a Gamma distribution with shape α and rate $\beta : \mu_i \sim \Gamma(\alpha, \beta)$. The gene mean is modified by the occurrence of outliers, scaled by observed and expected library size per cell and modified by simulated biological coefficients of variation (BCV). Then, counts are drawn from the Poisson distribution using the final means $\mu'_{i,j}$.

Splatter can be used to simulate the batch effect under two scenarios: a) the batch effect is promoted by differences in dropout rates; b)

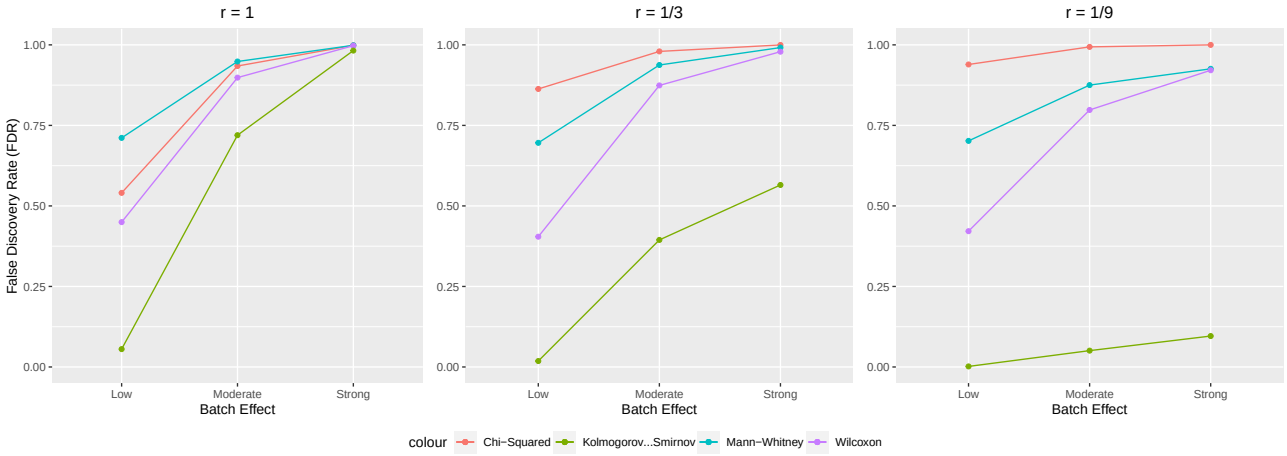

Figure 3: False Discovery Rate (FDR) for different batch effects and size ratios of batch 1 and batch 2.

the batch effect is promoted by additional noise. Here, we investigate the second scenario. In particular, we implement the batch effects as additional log-normal factor on gene means. The batch effect for a gene, b_i, is drawn from $LN(\mu, \sigma)$, where μ is the batch location factor and σ is the batch scale factor, and finally multiplied with the final gene mean $\mu'_{i,j}$.

We have varied both batch location factor and scale factor in order to simulate 3 data sets with different strengths of batch effects, defined as low, moderate and strong(Figure2). Each data set contains 500 cells (250 cells per batch) with 1000 genes each.

In addition, we varied the sample size of the two batches as in each simulation, we sampled 500 cells containing 1000 genes each, with the size ratio of the batches being

$$r \in \{1, 1/3, 1/4, 1/9, 1/19\}$$

This means that for instance when $r = 1$, we have equal number of cells from both batches, and when $r = 1/4$, we have 100 cells from batch 1, and we have 400 cells from batch 2.

5 Results

Finally, we evaluate the performance of the original kBET method (using a Chi-Square Test) with the three non-parametric tests that we introduced: Mann Whitney U-Test, Wilcoxon T-Test and Two-Sample K-S Test.

In order to compare the methods, we simulate three different datasets including two batches. We control for the batch effect (low, moderate, strong) and the size ratio of the two batches. The left panel of Figure 3 and Figure 4 show the simulated data with different strength of batch effect for a size ratio of 1/1 between the two batches.

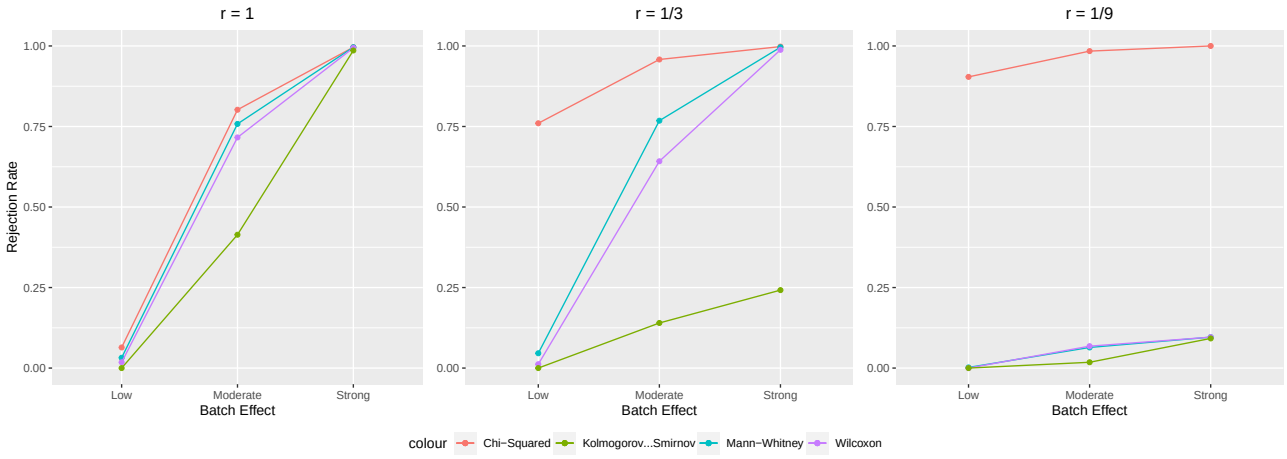

Figure 4: Rejection Rate for different batch effects and size ratios of batch 1 and batch 2.

For each of them, we compare the false discovery rate which is generated by subtracting the averaged p-values from 1 and the rejection rate which is the proportion of rejections. For the dataset with low batch effect, all tests yield a low rejection rate. In contrast, for the dataset with strong batch effect, samples were surrounded by samples from the same batch, thus all tests yield a high rejection rate. The test results are mostly consistent.

The rejection rate differs as the size ratio of batch 1 and batch 2 becomes smaller. For example, when the size ratio is 1/3 (the middle panel of Figure3 and Figure4), which means we have 125 cells from batch 1 and 375 cells from batch 2, although all tests are able to detect the different strength of batch effect, Chi-Square Test yields a relative high rejection rate when the batch effect is low while two sample K-S Test yields a relative low rejection rate when the batch effect is strong. When the size ratio is 1/9 (the right panel of Figure3 and Figure4), which means we have 50 cells from batch 1 and 450 cells from batch 2, the Chi-Square Test yields much higher rejection rates than the other tests.

6 Discussion

One of the important assumptions for these tests is whether the two samples are independent or not. The Mann Whitney U-Test and Two Sample K-S Test assume that two samples are independent, while the Wilcoxon T-Test assumes that they are dependent. Basically, the observed sample is randomly selected from the population and the observed sample and the expected sample can be regarded as two independent samples. However, on the other hand, the observed sample is still a subset of the population and there is a dependent relationship between them. More work has to be done to make a clear conclusion. Besides, the Wilcoxon T-Test assumes that the differences between two samples should follow normal distributions, so when the sample size is small, it is not applicable.

The neighborhood size k is an important factor to determine whether these tests are effective or not. Here k is assigned by the users. It may not work well if the user couldn't specify the value of k with high confidence. In addition, the assigned neighborhood size is not appropriate. It's worthy to discuss more about how to generate the optimal k automatically. One may consider obtaining the optimal size using cross validation. The github link containing all source codes for data simulation and the discussed test metrics can be found in the references.

References

[1] B. Phipson L. Zappia and A. Oshlack. Introduction to splatter. https://www.bioconductor.org/packages/release/bioc/vignettes/splatter/inst/doc/splatter.html, 2020.

[2] F. A. Wolf S. A. Teichmann M. B uttner, Z. Miao and F. J. Thei. "a test metric for assessing single-cell rna-seq batch correction. *Nature methods, vol. 16, no. 1, pp. 43–49*, 2019. https://pubmed.ncbi.nlm.nih.gov/30573817/.

[3] F. A. Wolf S. A. Teichmann M. B uttner, Z. Miao and F. J. Thei. "a test metric for assessing single-cell rna-seq batch correction - additional paper. *Nature methods, vol. 16, no. 1, pp. 43–49*, 2019. https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-018-0254-1/MediaObjects/41592_2018_254_MOESM1_ESM.pdf.

[4] G. Eraslani M. Buttner, Philipp A. kbet - k-nearest neighbour batch effect test. *Nature methods, vol. 16, no. 1, pp. 43–49*, 2019. https://github.com/theislab/kBET.

[5] XLSTAT. Which statistical test should you use? https://help.xlstat.com/s/article/which-statistical-test-should-you-use?language=en_US.

[6] A. Alvandi Y. Xu, F. Uzan. Nonparametric tests for batch effects detection. https://github.com/stat697BD/Batch-Effects-Detection, 2020.